



Figure 1: Examples of fast rates on functions without curvature. MetaGrad incurs logarithmic regret  $O(\ln T)$ , while AdaGrad incurs  $O(\sqrt{T})$  regret, matching its bound.

## A Extra Material Related to Section 3

In this section we gather extra material related to the fast rate examples from Section 3. We first provide simulations. Then we present the proofs of Theorems 2 and 3. And finally we give an example in which the unregularized hinge loss satisfies the Bernstein condition.

### A.1 Simulations: Logarithmic Regret without Curvature

We provide two simple simulation examples to illustrate the sufficient conditions from Theorems 2 and 3, and to show that such fast rates are not automatically obtained by previous methods for general functions. Both our examples are one-dimensional (so the full and diagonal algorithms coincide), and have a stable optimum (that good algorithms will converge to); yet the functions are based on absolute values, which are neither strongly convex nor smooth, so the gradient norms do not vanish near the optimum. As our baseline we include AdaGrad [9], because it is commonly used in practice [22, 27] and because, in the one-dimensional case, it implements GD with an adaptive tuning of the learning rate that is applicable to general convex functions.

In the first example, we consider offline convex optimization of the fixed function  $f_t(u) \equiv f(u) = |u - \frac{1}{4}|$ , which satisfies (3), because it is convex. In the second example, we look at stochastic optimization with convex functions  $f_t(u) = |u - x_t|$ , where the outcomes  $x_t = \pm \frac{1}{2}$  are chosen i.i.d. with probabilities 0.4 and 0.6. These probabilities satisfy (4) with  $\beta = 1$ . Their values are by no means essential, as long we avoid the worst case where the probabilities are equal.

Figure 1 graphs the results. We see that in both cases the regret of AdaGrad follows its  $O(\sqrt{T})$  bound, while MetaGrad achieves an  $O(\ln T)$  rate, as predicted by Theorems 2 and 3. This shows that MetaGrad achieves a type of adaptivity that is not achieved by AdaGrad. We should be careful in extending this conclusion to higher dimensions, though: whereas (the diagonal version of) AdaGrad uses a separate learning rate per dimension, MetaGrad shares learning rates between dimensions (unless we run a separate copy of MetaGrad per dimension, as suggested in the related work section).

### A.2 Proof of Theorem 2

*Proof.* By (3), applied with  $\mathbf{w} = \mathbf{w}_t$ , and Theorem 1, there exists a  $C > 0$  (depending on  $a$ ) such that, for all sufficiently large  $T$ ,

$$R_T^u \leq a\tilde{R}_T^u - bV_T^u \leq C\sqrt{V_T^u d \ln T} + Cd \ln T - bV_T^u$$

$$\leq \frac{\gamma}{2} C V_T^{\mathbf{u}} + \left( \frac{1}{2\gamma} + 1 \right) C d \ln T - b V_T^{\mathbf{u}} \quad \text{for all } \gamma > 0,$$

where the last inequality is based on  $\sqrt{xy} = \min_{\gamma>0} \frac{\gamma}{2}x + \frac{y}{2\gamma}$  for all  $x, y > 0$ . The result follows upon taking  $\gamma = \frac{2b}{C}$ .  $\square$

### A.3 Proof of Theorem 3

*Proof.* Abbreviate  $\tilde{r}_t^{\mathbf{u}} = (\mathbf{w}_t - \mathbf{u})^\top \mathbf{g}_t$ . Then, by (1), Jensen's inequality and the Bernstein condition, there exists a constant  $C > 0$  such that, for all sufficiently large  $T$ , the expected linearized regret is at most

$$\begin{aligned} \mathbb{E} [\tilde{R}_T^{\mathbf{u}^*}] &\leq C \mathbb{E} \left[ \sqrt{V_T^{\mathbf{u}^*}} d \ln T \right] + C d \ln T \leq C \sqrt{\mathbb{E} [V_T^{\mathbf{u}^*}] d \ln T} + C d \ln T \\ &\leq C \sqrt{B \sum_{t=1}^T (\mathbb{E} [\tilde{r}_t^{\mathbf{u}^*}])^\beta d \ln T} + C d \ln T. \end{aligned}$$

We will repeatedly use the fact that

$$x^\alpha y^{1-\alpha} = c_\alpha \inf_{\gamma>0} \left( \frac{x}{\gamma} + \gamma^{\frac{\alpha}{1-\alpha}} y \right) \quad \text{for any } x, y \geq 0 \text{ and } \alpha \in (0, 1), \quad (9)$$

where  $c_\alpha = (1-\alpha)^{1-\alpha} \alpha^\alpha$ . Applying this first with  $\alpha = 1/2$ ,  $x = B d \ln T$  and  $y = \sum_{t=1}^T (\mathbb{E} [\tilde{r}_t^{\mathbf{u}^*}])^\beta$ , we obtain

$$\sqrt{B \sum_{t=1}^T (\mathbb{E} [\tilde{r}_t^{\mathbf{u}^*}])^\beta d \ln T} \leq c_{1/2} \gamma_1 \sum_{t=1}^T (\mathbb{E} [\tilde{r}_t^{\mathbf{u}^*}])^\beta + \frac{c_{1/2}}{\gamma_1} B d \ln T \quad \text{for any } \gamma_1 > 0.$$

If  $\beta = 1$ , then  $\sum_{t=1}^T (\mathbb{E} [\tilde{r}_t^{\mathbf{u}^*}])^\beta = \mathbb{E} [\tilde{R}_T^{\mathbf{u}^*}]$  and the result follows by taking  $\gamma_1 = \frac{1}{2C c_{1/2}}$ . Alternatively, if  $\beta < 1$ , then we apply (9) a second time, with  $\alpha = \beta$ ,  $x = \mathbb{E} [\tilde{r}_t^{\mathbf{u}^*}]$  and  $y = 1$ , to find that, for any  $\gamma_2 > 0$ ,

$$\begin{aligned} \sqrt{B \sum_{t=1}^T (\mathbb{E} [\tilde{r}_t^{\mathbf{u}^*}])^\beta d \ln T} &\leq c_\beta c_{1/2} \gamma_1 \sum_{t=1}^T \left( \frac{\mathbb{E} [\tilde{r}_t^{\mathbf{u}^*}]}{\gamma_2} + \gamma_2^{\beta/(1-\beta)} \right) + \frac{c_{1/2}}{\gamma_1} B d \ln T \\ &= \frac{c_\beta c_{1/2} \gamma_1}{\gamma_2} \mathbb{E} [\tilde{R}_T^{\mathbf{u}^*}] + c_\beta c_{1/2} \gamma_1 \gamma_2^{\beta/(1-\beta)} T + \frac{c_{1/2}}{\gamma_1} B d \ln T. \end{aligned}$$

Taking  $\gamma_1 = \frac{\gamma_2}{2c_\beta c_{1/2} C}$ , this yields

$$\mathbb{E} [\tilde{R}_T^{\mathbf{u}^*}] \leq \gamma_2^{1/(1-\beta)} T + \frac{4C^2 c_{1/2}^2 c_\beta B d \ln T}{\gamma_2} + 2C d \ln T.$$

We may optimize over  $\gamma_2$  by a third application of (9), now with  $x = 4C^2 c_{1/2}^2 c_\beta B d \ln T$ ,  $y = T$  and  $\alpha = 1/(2-\beta)$ , such that  $\alpha/(1-\alpha) = 1/(1-\beta)$ :

$$\begin{aligned} \mathbb{E} [\tilde{R}_T^{\mathbf{u}^*}] &\leq \frac{1}{c_{1/(2-\beta)}} \left( 4C^2 c_{1/2}^2 c_\beta B d \ln T \right)^{1/(2-\beta)} T^{(1-\beta)/(2-\beta)} + 2C d \ln T \\ &= O \left( (B d \ln T)^{1/(2-\beta)} T^{(1-\beta)/(2-\beta)} + d \ln T \right), \end{aligned}$$

which completes the proof.  $\square$

### A.4 Unregularized Hinge Loss Example

As shown by Koolen, Grünwald, and Van Erven [19], the Bernstein condition is satisfied in the following classification task:

**Lemma 8** (Unregularized Hinge Loss Example). *Suppose that  $(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \dots$  are i.i.d. with  $Y_t$  taking values in  $\{-1, +1\}$ , and let  $f_t(\mathbf{u}) = \max\{0, 1 - Y_t \langle \mathbf{u}, \mathbf{X}_t \rangle\}$  be the hinge loss. Assume that both  $\mathcal{U}$  and the domain for  $\mathbf{X}_t$  are the  $d$ -dimensional unit ball. Then the  $(B, \beta)$ -Bernstein condition is satisfied with  $\beta = 1$  and  $B = \frac{2\lambda_{\max}}{\|\boldsymbol{\mu}\|}$ , where  $\lambda_{\max}$  is the maximum eigenvalue of  $\mathbb{E}[\mathbf{X}\mathbf{X}^\top]$  and  $\boldsymbol{\mu} = \mathbb{E}[Y\mathbf{X}]$ , provided that  $\|\boldsymbol{\mu}\| > 0$ .*

*In particular, if  $\mathbf{X}_t$  is uniformly distributed on the sphere and  $Y_t = \text{sign}(\langle \bar{\mathbf{u}}, \mathbf{X}_t \rangle)$  is the noiseless classification of  $\mathbf{X}_t$  according to the hyperplane with normal vector  $\bar{\mathbf{u}}$ , then  $B \leq \frac{c}{\sqrt{d}}$  for some absolute constant  $c > 0$ .*

Thus the version of the Bernstein condition that implies an  $O(d \ln T)$  rate is always satisfied for the hinge loss on the unit ball, except when  $\|\boldsymbol{\mu}\| = 0$ , which is very natural to exclude, because it implies that the expected hinge loss is 1 (its maximal value) for all  $\mathbf{u}$ , so there is nothing to learn. It is common to add  $\ell_2$ -regularization to the hinge loss to make it strongly convex, but this example shows that that is not necessary to get logarithmic regret.

## B Master Regret Bound (Proof of Lemma 4)

*Proof.* To prove Lemma 4, we start by bounding  $e^{-\alpha \ell_t^\eta(\mathbf{w}_t^\eta)}$  by its tangent at  $\mathbf{w}_t^\eta = \mathbf{w}_t$ :

$$e^{-\alpha \ell_t^\eta(\mathbf{w}_t^\eta)} \leq 1 + \alpha \eta (\mathbf{w}_t - \mathbf{w}_t^\eta)^\top \mathbf{g}_t \quad \text{for any } \eta \in (0, \frac{2}{3DG}]. \quad (10)$$

For the full case, where  $\alpha = \alpha^{\text{full}} = 1$ , this follows directly from the “prod bound”  $e^{x-x^2} \leq 1 + x$  with  $x = \eta (\mathbf{w}_t - \mathbf{w}_t^\eta)^\top \mathbf{g}_t$ , which has previously been used in the prediction with expert advice setting [3, 11, 16] and holds for any  $x \geq -2/3$ . In the diagonal case, (10) does not hold with  $\alpha = 1$ , but it can be proved with  $\alpha = \alpha^{\text{diag}} = 1/d$  by an application of Jensen’s inequality combined with a separate prod bound per dimension:

$$\begin{aligned} e^{-\alpha \ell_t^\eta(\mathbf{w}_t^\eta)} &= e^{\sum_i \frac{1}{d} (\eta(w_{t,i} - w_{t,i}^\eta) g_{t,i} - \eta^2 (w_{t,i} - w_{t,i}^\eta)^2 g_{t,i}^2)} \stackrel{\text{Jensen}}{\leq} \sum_i \frac{1}{d} e^{(\eta(w_{t,i} - w_{t,i}^\eta) g_{t,i} - \eta^2 (w_{t,i} - w_{t,i}^\eta)^2 g_{t,i}^2)} \\ &\stackrel{\text{prod bound}}{\leq} \sum_i \frac{1}{d} (1 + \eta(w_{t,i} - w_{t,i}^\eta) g_{t,i}) = 1 + \alpha \eta (\mathbf{w}_t - \mathbf{w}_t^\eta)^\top \mathbf{g}_t. \end{aligned}$$

We proceed to show that the potential  $\Phi_T$  is non-increasing:

$$\begin{aligned} \Phi_{T+1} - \Phi_T &= \sum_{\eta} \pi_1^\eta e^{-\alpha \sum_{t=1}^T \ell_t^\eta(\mathbf{w}_t^\eta)} \left( e^{-\alpha \ell_T^\eta(\mathbf{w}_{T+1}^\eta)} - 1 \right) \\ &\leq \sum_{\eta} \pi_1^\eta e^{-\alpha \sum_{t=1}^T \ell_t^\eta(\mathbf{w}_t^\eta)} \alpha \eta (\mathbf{w}_{T+1} - \mathbf{w}_{T+1}^\eta)^\top \mathbf{g}_{T+1} = 0, \end{aligned}$$

where the inequality is the tangent bound (10), and the final equality is by definition of the master prediction (in fact, it can be taken as the motivation for the master’s definition)

$$\mathbf{w}_{T+1} = \frac{\sum_{\eta} \pi_{T+1}^\eta \eta \mathbf{w}_{T+1}^\eta}{\sum_{\eta} \pi_{T+1}^\eta \eta} = \frac{\sum_{\eta} \pi_1^\eta e^{-\alpha \sum_{t=1}^T \ell_t^\eta(\mathbf{w}_t^\eta)} \eta \mathbf{w}_{T+1}^\eta}{\sum_{\eta} \pi_1^\eta e^{-\alpha \sum_{t=1}^T \ell_t^\eta(\mathbf{w}_t^\eta)} \eta}.$$

Since  $\Phi_0 = 1$  is trivial, this completes the proof of the lemma.  $\square$

## C Slave Regret Bound (Proof of Lemma 5)

*Proof.* For any distributions  $P$  and  $Q$  on  $\mathbb{R}^d$ , let  $\text{KL}(P\|Q) = \mathbb{E}_P[\ln \frac{dP}{dQ}]$  denote the Kullback-Leibler divergence of  $P$  from  $Q$ , and let  $\boldsymbol{\mu}_P = \mathbb{E}_P[\mathbf{u}]$  denote the mean of  $P$ . In addition, let  $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  denote a normal distribution with mean  $\boldsymbol{\mu}$  and covariance matrix  $\boldsymbol{\Sigma}$ .

In round  $t$ , we play according to the mean of a multivariate Gaussian distribution  $P_t$ . In the first round, this is a normal distribution, which plays the role of a prior:

$$P_1 = \mathcal{N}(\mathbf{0}, D^2 \mathbf{I}).$$

Then we update using the exponential weights update, followed by a projection onto  $\mathcal{P} = \{P : \mu_P \in \mathcal{U}\}$ , such that the mean stays in the allowed domain  $\mathcal{U}$ :

$$d\tilde{P}_{t+1}(\mathbf{u}) = \frac{e^{-\ell_t^\eta(\mathbf{u})} dP_t(\mathbf{u})}{\int_{\mathbb{R}^d} e^{-\ell_t^\eta(\mathbf{u}') } dP_t(\mathbf{u}')}, \quad P_{t+1} = \arg \min_{P \in \mathcal{P}} \text{KL}(P \| \tilde{P}_t).$$

To see that Algorithm 2 implements this algorithm, we prove by induction that

$$P_t = \mathcal{N}(\mathbf{w}_t^\eta, \Sigma_t^\eta).$$

For  $t = 1$  this is clear, and if it holds for any  $t$  then it can be verified by comparing densities that  $\tilde{P}_{t+1} = \mathcal{N}(\tilde{\mathbf{w}}_{t+1}^\eta, \Sigma_{t+1}^\eta)$ . Since it is well-known that the KL-projection of a Gaussian  $\mathcal{N}(\mu, \Sigma)$  onto  $\mathcal{P}$  is another Gaussian  $\mathcal{N}(\nu, \Sigma)$  with the same covariance matrix and mean  $\nu \in \mathcal{U}$  that minimizes  $\frac{1}{2}(\nu - \mu)^\top \Sigma^{-1}(\nu - \mu)$ , it then follows that  $P_{t+1} = \mathcal{N}(\mathbf{w}_{t+1}^\eta, \Sigma_{t+1}^\eta)$ . For completeness we provide a proof of this last result in Lemma 9 of Appendix F.

It now remains to bound the regret. Since  $\mathcal{P}$  is convex, the Pythagorean inequality for Kullback-Leibler divergence implies that

$$\text{KL}(Q \| \tilde{P}_{t+1}) \geq \text{KL}(Q \| P_{t+1}) + \text{KL}(P_{t+1} \| \tilde{P}_{t+1}) \geq \text{KL}(Q \| P_{t+1})$$

for all  $Q \in \mathcal{P}$ . The following telescoping sum therefore gives us that

$$\begin{aligned} \text{KL}(Q \| P_1) &\geq \sum_{t=1}^T \text{KL}(Q \| P_t) - \text{KL}(Q \| P_{t+1}) \geq \sum_{t=1}^T \text{KL}(Q \| P_t) - \text{KL}(Q \| \tilde{P}_{t+1}) \\ &= \sum_{t=1}^T -\ln \mathbb{E}_{P_t} [e^{-\ell_t^\eta(\mathbf{u})}] - \mathbb{E}_Q [\ell_t^\eta(\mathbf{u})]. \end{aligned} \quad (11)$$

This may be interpreted as a regret bound in the space of distributions, which we will now relate to our regret of interest. If  $M_t = M_t^{\text{full}}$ , then Lemma 10 in Appendix G implies that

$$-\ln \mathbb{E}_{P_t} [e^{-\ell_t^\eta(\mathbf{u})}] \geq \ell_t^\eta(\mathbf{w}_t^\eta)$$

because  $\mathbf{w}_t^\eta$  is the mean of  $P_t$ . Alternatively, if  $M_t = M_t^{\text{diag}}$  then  $P_t$  has diagonal covariance  $\Sigma_t^\eta$ , and we can use Lemma 10 again to draw the same conclusion.

To control  $\mathbb{E}_Q [\ell_t^\eta(\mathbf{u})]$ , we may restrict attention (without loss of generality by a standard maximum entropy argument) to normal distributions  $Q = \mathcal{N}(\mu, D^2 \Sigma)$  with mean  $\mu \in \mathcal{U}$  and covariance  $\Sigma \succ \mathbf{0}$  (expressed relative to the prior variance  $D^2$ ). Then, using the cyclic property and linearity of the trace,

$$\begin{aligned} \mathbb{E}_Q [\ell_t^\eta(\mathbf{u})] &= -\eta(\mathbf{w}_t - \mu)^\top \mathbf{g}_t + \eta^2(\mathbf{w}_t^\top M_t \mathbf{w}_t - 2\mu^\top M_t \mathbf{w}_t + \mathbb{E}_Q [\text{tr}(\mathbf{u}^\top M_t \mathbf{u})]) \\ &= -\eta(\mathbf{w}_t - \mu)^\top \mathbf{g}_t + \eta^2(\mathbf{w}_t^\top M_t \mathbf{w}_t - 2\mu^\top M_t \mathbf{w}_t + \text{tr}(\mathbb{E}_Q [\mathbf{u} \mathbf{u}^\top] M_t)) \\ &= -\eta(\mathbf{w}_t - \mu)^\top \mathbf{g}_t + \eta^2(\mathbf{w}_t^\top M_t \mathbf{w}_t - 2\mu^\top M_t \mathbf{w}_t + \text{tr}((D^2 \Sigma + \mu \mu^\top) M_t)) \\ &= -\eta(\mathbf{w}_t - \mu)^\top \mathbf{g}_t + \eta^2((\mu - \mathbf{w}_t)^\top M_t (\mu - \mathbf{w}_t) + D^2 \text{tr}(\Sigma M_t)) \\ &= \ell_t^\eta(\mu) + \eta^2 D^2 \text{tr}(\Sigma M_t). \end{aligned}$$

Finally, it remains to work out

$$\text{KL}(Q \| P_1) = \frac{1}{2D^2} \|\mu\|^2 + \frac{1}{2} (-\ln \det \Sigma + \text{tr}(\Sigma) - d).$$

We have now bounded all the pieces in (11). Putting them all together with the choice  $\mu = \mathbf{u}$  and optimizing the bound in  $\Sigma$  gives:

$$\begin{aligned} \sum_{t=1}^T \ell_t^\eta(\mathbf{w}_t^\eta) - \sum_{t=1}^T \ell_t^\eta(\mathbf{u}) &\leq \frac{1}{2D^2} \|\mathbf{u}\|^2 + \frac{1}{2} \inf_{\Sigma \succ \mathbf{0}} \left\{ -\ln \det \Sigma + \text{tr} \left( \Sigma \left( \mathbf{I} + 2\eta^2 D^2 \sum_{t=1}^T M_t \right) \right) - d \right\} \\ &= \frac{1}{2D^2} \|\mathbf{u}\|^2 + \frac{1}{2} \ln \det \left( \mathbf{I} + 2\eta^2 D^2 \sum_{t=1}^T M_t \right), \end{aligned} \quad (12)$$

where the minimum is attained at  $\Sigma = \left( \mathbf{I} + 2\eta^2 D^2 \sum_{t=1}^T M_t \right)^{-1}$ .  $\square$

## D Composition Proofs

Throughout this section we abbreviate  $\mathbf{S}_T = \sum_{t=1}^T \mathbf{M}_t$ .

### D.1 Proof of Theorem 6

*Proof.* We start with

$$\begin{aligned} 0 &\stackrel{\text{Lemma 4}}{\geq} \frac{1}{\alpha} \ln \Phi_T \geq \frac{1}{\alpha} \ln \pi_1^\eta - \sum_{t=1}^T \ell_t^\eta(\mathbf{w}_t^\eta) \\ &\stackrel{\text{Lemma 5}}{\geq} \frac{1}{\alpha} \ln \pi_1^\eta - \sum_{t=1}^T \ell_t^\eta(\mathbf{u}) - \frac{1}{2D^2} \|\mathbf{u}\|^2 - \frac{1}{2} \ln \det(\mathbf{I} + 2\eta^2 D^2 \mathbf{S}_T). \end{aligned}$$

Now expanding the definition (6) of the surrogate losses we find

$$\eta \sum_{t=1}^T (\mathbf{w}_t - \mathbf{u})^\top \mathbf{g}_t \leq \frac{1}{2D^2} \|\mathbf{u}\|^2 - \frac{1}{\alpha} \ln \pi_1^\eta + \eta^2 V_T^\mathbf{u} + \frac{1}{2} \ln \det(\mathbf{I} + 2\eta^2 D^2 \mathbf{S}_T),$$

in which we may divide by  $\eta$  to obtain the claim.  $\square$

### D.2 Proof of Theorem 7

*Proof.* In principle we would like to directly select the  $\eta$  that optimizes the regret bound from Theorem 6. But unfortunately we cannot tractably minimize that bound, since  $\eta$  occurs in the  $\ln \det$ . To bring the  $\eta$  out, we apply the variational form from (12) to Theorem 6 to obtain

$$\tilde{R}_T^\mathbf{u} \leq \inf_{\Sigma \succ \mathbf{0}} \eta_i (V_T^\mathbf{u} + D^2 \text{tr}(\Sigma \mathbf{S}_T)) + \frac{\frac{1}{D^2} \|\mathbf{u}\|^2 - \frac{2}{\alpha} \ln \pi_1^{\eta_i} - \ln \det(\Sigma) + \text{tr}(\Sigma) - d}{2\eta_i} \quad (13)$$

for all grid points  $\eta_i$ . This leads to an upper bound by plugging in a near-optimal choice for  $\Sigma$ , which we choose as

$$\begin{array}{cc} \text{full} & \text{diag} \\ \Sigma = (\mathbf{I} + c\mathbf{S}_T)^{-1} & \Sigma = \frac{1}{D^2} \text{diag}(V_{T,1}^\mathbf{u}, \dots, V_{T,d}^\mathbf{u}) \mathbf{S}_T^{-1}, \end{array}$$

where  $c := \text{rk}(\mathbf{S}_T) \left( \frac{D^2}{V_T^\mathbf{u}} - \frac{1}{\text{tr}(\mathbf{S}_T)} \right)$  is non-negative because  $V_T^\mathbf{u} \leq D^2 \sum_{t=1}^T \|\mathbf{g}_t\|^2 = D^2 \text{tr}(\mathbf{S}_T)$  by Cauchy-Schwarz. We proceed to bound terms involving  $\Sigma$  above. In the diagonal case, we use that  $D^2 \text{tr}(\Sigma \mathbf{S}_T) = V_T^\mathbf{u}$  and  $\Sigma \prec \mathbf{I}$  because  $V_{T,i}^\mathbf{u} \leq D^2 \sum_{t=1}^T g_{t,i}^2$  by Cauchy-Schwarz. In the full case, we also have  $\Sigma \prec \mathbf{I}$ . In addition, we observe that  $\mathbf{S}_T$  and  $\Sigma$  share the same eigenbasis, so we may work in that basis. As  $\Sigma \mathbf{S}_T$  has  $\text{rk}(\mathbf{S}_T)$  non-zero eigenvalues, we may pull out a factor  $\text{rk}(\mathbf{S}_T)$  and replace the trace by a uniform average of the eigenvalues. Then Jensen's inequality for the concave function  $x \mapsto \frac{x}{1+cx}$  for  $x \geq 0$  gives

$$D^2 \text{tr}(\Sigma \mathbf{S}_T) \stackrel{\text{Jensen}}{\leq} \frac{D^2 \text{tr}(\mathbf{S}_T)}{\left(1 + \frac{c}{\text{rk}(\mathbf{S}_T)} \text{tr}(\mathbf{S}_T)\right)} = V_T^\mathbf{u}.$$

Thus, in both cases we have  $D^2 \text{tr}(\Sigma \mathbf{S}_T) \leq V_T^\mathbf{u}$  and  $\Sigma \prec \mathbf{I}$ , which implies that  $\text{tr}(\Sigma) \leq \text{tr}(\mathbf{I}) = d$  and that

$$\Xi_T := -\ln \det \Sigma \geq 0.$$

Finally, by construction of the grid, for any  $\eta \in [\frac{1}{5DG\sqrt{T}}, \frac{2}{5DG}]$  there exists a grid point  $\eta_i \in [\frac{\eta}{2}, \eta]$ , and the prior costs of this grid point satisfy

$$-\ln \pi_1^{\eta_i} \leq 2 \ln(2 + i) \leq 2 \ln\left(3 + \frac{1}{2} \log_2 T\right).$$

Plugging these bounds into (13) and abbreviating

$$A := \frac{1}{D^2} \|\mathbf{u}\|^2 + \frac{4}{\alpha} \ln\left(3 + \frac{1}{2} \log_2 T\right) + \Xi_T \geq 4 \ln 3,$$

we obtain

$$\tilde{R}_T^u \leq 2\eta V_T^u + \frac{A}{\eta}.$$

Subsequently tuning  $\eta$  optimally as

$$\hat{\eta} = \sqrt{\frac{A}{2V_T^u}} \geq \frac{\sqrt{2\ln 3}}{DG\sqrt{T}} \geq \frac{1}{5DG\sqrt{T}}$$

is allowed when  $\hat{\eta} \leq \frac{2}{5DG}$ , and gives  $\tilde{R}_T^u \leq \sqrt{8V_T^u A}$ . Alternatively, if  $\hat{\eta} \geq \frac{2}{5DG}$ , then we plug in  $\eta = \frac{2}{5DG}$  and obtain  $\tilde{R}_T^u \leq \frac{4}{5DG} V_T^u + \frac{5}{2} DGA \leq 5DGA$ , where the second inequality follows from the constraint on  $\hat{\eta}$ . In both cases, we find that

$$\tilde{R}_T^u \leq \sqrt{8V_T^u A} + 5DGA,$$

which results in the first claim of the theorem upon observing that, for the full version of the algorithm,  $\Xi_T \leq \text{rk}(\mathbf{S}_T) \ln \left( \frac{D^2 \text{tr}(\mathbf{S}_T)}{V_T^u} \right)$  by Jensen's inequality and  $\Xi_T \leq \ln \det \left( \mathbf{I} + \frac{D^2 \text{rk}(\mathbf{S}_T)}{V_T^u} \mathbf{S}_T \right)$  by monotonicity of  $\ln \det$ .

To prove the second claim, we instead take the comparator covariance  $\Sigma = \mathbf{I}$  equal to the prior covariance and again use  $V_T^u \leq D^2 \text{tr}(\mathbf{S}_T)$  to find

$$\begin{aligned} \tilde{R}_T^u &\leq \eta_i (V_T^u + D^2 \text{tr}(\mathbf{S}_T)) + \frac{\frac{1}{D^2} \|\mathbf{u}\|^2 - \frac{2}{\alpha} \ln \pi_1^{\eta_i}}{2\eta_i} \leq 2\eta_i D^2 \text{tr}(\mathbf{S}_T) + \frac{\frac{1}{D^2} \|\mathbf{u}\|^2 - \frac{2}{\alpha} \ln \pi_1^{\eta_i}}{2\eta_i} \\ &\leq 2\eta D^2 \text{tr}(\mathbf{S}_T) + \frac{\frac{1}{D^2} \|\mathbf{u}\|^2 + \frac{4}{\alpha} \ln(3 + \frac{1}{2} \log_2 T)}{\eta} \end{aligned}$$

for all  $\eta \in [\frac{1}{5DG\sqrt{T}}, \frac{2}{5DG}]$ . Tuning  $\eta$  as

$$\hat{\eta} = \sqrt{\frac{\frac{1}{D^2} \|\mathbf{u}\|^2 + \frac{4}{\alpha} \ln(3 + \frac{1}{2} \log_2 T)}{2D^2 \text{tr}(\mathbf{S}_T)}} \geq \sqrt{\frac{4\ln 3}{2D^2 G^2 T}} \geq \frac{1}{5DG\sqrt{T}}$$

is allowed when  $\hat{\eta} \leq \frac{2}{5DG}$ , and gives

$$\tilde{R}_T^u \leq \sqrt{8D^2 \text{tr}(\mathbf{S}_T) \left( \frac{1}{D^2} \|\mathbf{u}\|^2 + \frac{4}{\alpha} \ln(3 + \frac{1}{2} \log_2 T) \right)}.$$

Alternatively, if  $\hat{\eta} \geq \frac{2}{5DG}$ , then we plug in  $\eta = \frac{2}{5DG}$  and obtain

$$\begin{aligned} \tilde{R}_T^u &\leq \frac{4}{5DG} D^2 \text{tr}(\mathbf{S}_T) + \frac{5}{2} DG \left( \frac{1}{D^2} \|\mathbf{u}\|^2 + \frac{4}{\alpha} \ln(3 + \frac{1}{2} \log_2 T) \right) \\ &\leq 5DG \left( \frac{1}{D^2} \|\mathbf{u}\|^2 + \frac{4}{\alpha} \ln(3 + \frac{1}{2} \log_2 T) \right), \end{aligned}$$

where the second inequality follows from the constraint on  $\hat{\eta}$ . In both cases, the second claim of the theorem follows.  $\square$

## E Discussion of the Choice of Grid Points and Prior Weights

We now think about the choice of the grid and corresponding prior. Theorem 6 above implies that any two  $\eta$  that are within a constant factor of each other will guarantee the same bound up to a constant factor. Since  $\eta$  is a continuous parameter, this suggests choosing a prior that is approximately uniform for  $\ln \eta$ , which means it should have a density that looks like  $1/\eta$ . Although Theorem 6 does not show it, there is never any harm in taking too many grid points, because grid points that are very close together will behave as a single point with combined prior mass. If we disregard computation, we would therefore like to use the prior discussed by Chernov and Vovk [4], which is very close to uniform on  $\ln \eta$  and has density

$$\pi(\eta) = \frac{C}{\eta \log_2^2(\frac{5}{2} DG\eta)},$$

where we include the factor  $\frac{5}{2}DG$  to make the prior invariant under rescalings of the problem, and  $C$  is a normalizing constant that makes the prior integrate to 1. To adapt this prior to a discrete grid, we need to integrate this density between grid points and assign prior masses:

$$\pi_1^{\eta_i} := \int_{\eta_{i+1}}^{\eta_i} \pi(\eta) d\eta = \frac{C \ln(2)}{-\log_2(\frac{5}{2}DG\eta)} \Big|_{\eta_{i+1}}^{\eta_i}.$$

For the exponentially spaced grid in (8), this evaluates to the prior weights  $\pi_1^{\eta_i}$  specified there.

## F Projection of Gaussians

It is well-known that the projection of a Gaussian onto the set of distributions with mean in the convex set  $\mathcal{U}$  is also a Gaussian with the same covariance matrix. This result follows easily from, for instance, Theorems 1.8.5 and 1.8.2 of Ihara [15], but we include a short proof for completeness:

**Lemma 9.** *Let  $\tilde{P}_t = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  be Gaussian and let  $P_t = \arg \min_{P: \boldsymbol{\mu}_P \in \mathcal{U}} \text{KL}(P \| \tilde{P}_t)$  be its projection onto the set of distributions with mean in  $\mathcal{U}$ . Then  $P_t$  is also Gaussian with the same covariance matrix:*

$$P_t = \mathcal{N}(\boldsymbol{\nu}, \boldsymbol{\Sigma})$$

for  $\boldsymbol{\nu} \in \mathcal{U}$  that minimizes  $\frac{1}{2}(\boldsymbol{\nu} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\nu} - \boldsymbol{\mu})$ .

*Proof.* Let  $P$  be an arbitrary distribution with mean  $\boldsymbol{\nu} \in \mathcal{U}$ , and let  $R = \mathcal{N}(\boldsymbol{\nu}, \boldsymbol{\Sigma})$ . Then by straight-forward algebra and nonnegativity of Kullback-Leibler divergence it can be verified that

$$\text{KL}(P \| \tilde{P}_t) = \text{KL}(P \| R) + \text{KL}(R \| \tilde{P}_t) \geq \text{KL}(R \| \tilde{P}_t).$$

Thus the minimum over all  $P$  is achieved by a Gaussian with the same covariance matrix as  $\tilde{P}_t$ . It remains to find the mean of the projection, which is the  $\boldsymbol{\nu} \in \mathcal{U}$  that minimizes

$$\text{KL}(R \| \tilde{P}_t) = \frac{1}{2}(\boldsymbol{\nu} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\nu} - \boldsymbol{\mu}),$$

as required.  $\square$

## G Gaussian Exp-concavity

Exp-concavity of  $\ell_t^\eta(\mathbf{u})$  means that  $\mathbb{E} \left[ e^{-\ell_t^\eta(\mathbf{u})} \right] \leq e^{-\ell_t^\eta(\boldsymbol{\mu})}$  for any distribution on  $\mathbf{u} \in \mathbb{R}^d$ . Although this does not hold for general distributions with support outside of  $\mathcal{U}$ , it does hold if we restrict attention to certain types of Gaussians:

**Lemma 10** (Gaussian exp-concavity). *Let  $0 < \eta \leq \frac{1}{5DG}$ . Consider a Gaussian distribution with mean  $\boldsymbol{\mu} \in \mathcal{U}$  and arbitrary covariance  $\boldsymbol{\Sigma} \succ \mathbf{0}$  in the full case or diagonal  $\boldsymbol{\Sigma} \succ \mathbf{0}$  in the diagonal case. Then*

$$\mathbb{E}_{\mathbf{u} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})} \left[ e^{-\ell_t^\eta(\mathbf{u})} \right] \leq e^{-\ell_t^\eta(\boldsymbol{\mu})}$$

*Proof.* We first consider the full case. Abbreviating  $r := (\mathbf{w}_t - \boldsymbol{\mu})^\top \mathbf{g}_t$  and  $s := (\boldsymbol{\mu} - \mathbf{u})^\top \mathbf{g}_t$ , from the definition (6) of  $\ell_t^\eta$  we get

$$\begin{aligned} \ell_t^\eta(\boldsymbol{\mu}) - \ell_t^\eta(\mathbf{u}) &= \eta(\boldsymbol{\mu} - \mathbf{u})^\top \mathbf{g}_t - \eta^2 (2(\boldsymbol{\mu} - \mathbf{w}_t)^\top \mathbf{g}_t \mathbf{g}_t^\top (\boldsymbol{\mu} - \mathbf{u}) + (\boldsymbol{\mu} - \mathbf{u})^\top \mathbf{g}_t \mathbf{g}_t^\top (\boldsymbol{\mu} - \mathbf{u})) \\ &= \eta s - \eta^2 (2rs + s^2). \end{aligned}$$

Since  $\mathbf{u} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  implies  $s \sim \mathcal{N}(0, v)$  with  $v = \mathbf{g}_t^\top \boldsymbol{\Sigma} \mathbf{g}_t$ , the claim collapses to

$$1 \geq \mathbb{E}_{\mathbf{u} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})} \left[ e^{\ell_t^\eta(\boldsymbol{\mu}) - \ell_t^\eta(\mathbf{u})} \right] = \mathbb{E}_{s \sim \mathcal{N}(0, v)} \left[ e^{\eta s - \eta^2 (2rs + s^2)} \right] = \frac{e^{\frac{\eta^2 v (1 - 2\eta r)^2}{2(1 + 2\eta^2 v)}}}{\sqrt{1 + 2\eta^2 v}},$$

which is equivalent to

$$(1 - 2\eta r)^2 \eta^2 v \leq (1 + 2\eta^2 v) \ln(1 + 2\eta^2 v).$$

The left-hand side is maximized over  $r \in [-D^{\text{full}}G^{\text{full}}, D^{\text{full}}G^{\text{full}}]$  at  $r = -D^{\text{full}}G^{\text{full}}$ . So it suffices to establish

$$v(1 + 2\eta D^{\text{full}}G^{\text{full}})^2 \eta^2 \leq (1 + 2\eta^2 v) \ln(1 + 2\eta^2 v).$$

Now the right-hand is convex in  $v$  and hence bounded below by its tangent at  $v = 0$ , which is  $2\eta^2 v$ . The proof is completed by observing that  $(1 + 2\eta D^{\text{full}}G^{\text{full}})^2 \leq 2$  by the assumed bound on  $\eta$ .

It remains to consider the diagonal case. There the surrogate loss (6) is a sum over dimensions, say  $\ell_t^\eta(\mathbf{u}) = \sum_{i=1}^d \ell_{t,i}^\eta(u_i)$ . For a Gaussian with diagonal covariance matrix  $\Sigma$  the coordinates of  $\mathbf{u}$  are independent, and hence

$$\mathbb{E}_{\mathbf{u} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)} \left[ e^{-\ell_t^\eta(\mathbf{u})} \right] = \prod_{i=1}^d \mathbb{E}_{u_i \sim \mathcal{N}(\mu_i, \Sigma_{i,i})} \left[ e^{-\ell_{t,i}^\eta(u_i)} \right] \leq \prod_{i=1}^d e^{-\ell_{t,i}^\eta(\mu_i)} = e^{-\ell_t^\eta(\boldsymbol{\mu})},$$

where the inequality is the result for the full case applied to each dimension separately.  $\square$

## H Bernstein for Linearized Excess Loss

Let  $f : \mathcal{U} \rightarrow \mathbb{R}$  be a convex function drawn from distribution  $\mathbb{P}$  with stochastic optimum  $\mathbf{u}^* = \arg \min_{\mathbf{u} \in \mathcal{U}} \mathbb{E}_{f \sim \mathbb{P}}[f(\mathbf{u})]$ . For any  $\mathbf{w} \in \mathcal{U}$ , we now show that the Bernstein condition for the excess loss  $X := f(\mathbf{w}) - f(\mathbf{u}^*)$  implies the Bernstein condition with the same exponent  $\beta$  for the linearized excess loss  $Y := (\mathbf{w} - \mathbf{u}^*)^\top \nabla f(\mathbf{w})$ . These variables satisfy  $Y \geq X$  by convexity of  $f$  and  $Y \leq C := D^{\text{full}}G^{\text{full}}$ .

**Lemma 11.** *For  $\beta \in (0, 1]$ , let  $X$  be a  $(B, \beta)$ -Bernstein random variable:*

$$\mathbb{E}[X^2] \leq B \mathbb{E}[X]^\beta.$$

*Then any bounded random variable  $Y \leq C$  with  $Y \geq X$  pointwise satisfies the  $(B', \beta)$ -Bernstein condition*

$$\mathbb{E}[Y^2] \leq B' \mathbb{E}[Y]^\beta$$

*for  $B' = \max \left\{ B, \frac{2}{\beta} C^{2-\beta} \right\}$ .*

*Proof.* For  $\beta \in (0, 1)$  we will use the fact that

$$z^\beta = c_\beta \inf_{\gamma > 0} \left( \frac{z}{\gamma} + \gamma^{\frac{\beta}{1-\beta}} \right) \quad \text{for any } z \geq 0,$$

with  $c_\beta = (1 - \beta)^{1-\beta} \beta^\beta$ . For  $\gamma = \left( \frac{1-\beta}{\beta} \mathbb{E}[Y] \right)^{1-\beta}$  we therefore have

$$\begin{aligned} \mathbb{E}[X^2] - B' \mathbb{E}[X]^\beta &\geq \mathbb{E}[X^2] - B' c_\beta \left( \frac{\mathbb{E}[X]}{\gamma} + \gamma^{\frac{\beta}{1-\beta}} \right) \geq \mathbb{E}[Y^2] - B' c_\beta \left( \frac{\mathbb{E}[Y]}{\gamma} + \gamma^{\frac{\beta}{1-\beta}} \right) \\ &= \mathbb{E}[Y^2] - B' \mathbb{E}[Y]^\beta, \end{aligned} \tag{14}$$

where the second inequality holds because  $x^2 - c_\beta B' x / \gamma$  is a decreasing function of  $x \leq C$  for  $\gamma \leq \frac{c_\beta B'}{2C}$ , which is satisfied by the choice of  $B'$ . This proves the lemma for  $\beta \in (0, 1)$ . The claim for  $\beta = 1$  follows by taking the limit  $\beta \rightarrow 1$  in (14).  $\square$